

# 法令英訳のための対訳辞書の構築とその利用

小川 泰弘<sup>†</sup> 今井 一裕<sup>†</sup> 外山 勝彦<sup>†</sup> 松浦 好治<sup>‡</sup>

<sup>†</sup>名古屋大学大学院情報科学研究科 <sup>‡</sup>名古屋大学大学院法学研究科

yasuhiro@is.nagoya-u.ac.jp

## 1 はじめに

現在、国際取引の円滑化や対日投資の促進、法情報の国際的共有のための基盤構築、法整備支援の充実などの要求から、我が国の法令を翻訳する社会的必要性が高まっている [1]。2005 年 1 月、政府は法令外国語訳・実施推進検討会議を設置し、法令外国語訳推進のための基盤整備に関する検討を開始した。筆者らはそれに参画し、計算機を利用した法令翻訳支援に取り組んでいる。

本稿では、現在の法令翻訳における問題点と、それを自然言語処理の立場からどのように解決するかについて概説する。具体的には、標準対訳辞書の構築や、その辞書に基づく訳語統一の支援について述べる。

## 2 法令翻訳の問題点

従来の法令翻訳は、関係省庁や民間企業などにより、人手で個別的・断片的に行われてきた。そのため、ある法令とその関連法令のそれぞれの英訳を入手しても、同一の用語に対して異なる訳が与えられている場合があり、法令の内容を理解するのが容易ではなかったり、混乱を生じたりすることがあった。また、意味内容や文脈によっては、一つの用語に対して複数の訳語を使い分ける必要があるが、その基準も不明瞭であった。

そうした問題に対しては、法令文書特有の専門用語や定型表現に関する標準対訳辞書を構築し、訳語に加えて用例や使い分け基準も収録することが考えられる。そのため、法令外国語訳・実施推進検討会議において標準対訳辞書の構築が提言されている。

我々は、自然言語処理技術を応用し、この標準対訳辞書の構築を支援した。また、標準対訳辞書の完成後には、辞書に準拠して翻訳する必要があるが、そのための支援システムも開発している。

## 3 対訳表現の獲得

本研究では、これまで個別に翻訳されてきた法令の英訳を収集し、法令英訳に使用する標準対訳辞書を構

築するために、計算機を利用し、以下の方法により対訳表現を獲得した。

### 3.1 法令対訳コーパスの構築

15 府省等から、これまで個別に翻訳してきた法令 171 本の提供を受け、人手により和文・英文間の対応を付け、対訳コーパスを構築した。通常の対訳コーパスにおいて文ごとの対応を付ける際には、1 文が複数の文に翻訳される場合などを考慮する必要がある。しかし、幸いなことに、法令文には条や項などの単位があり、その構造は翻訳文でも保持される。よって、そうした条・項の単位ごとの対応付けすることによって、1 文が 2 文に翻訳されている場合などについても簡単に対応を付けることができた。

ただし、提出された法令翻訳の中には紙面で提出されたものがあつた。OCR による処理などに時間が掛かったため、そうしたものを除く法令 153 本 (対応文 39,560 組) を用いて、以下の対訳自動抽出を行った。

### 3.2 対訳表現の自動抽出

対訳表現の自動抽出に関しては、様々な手法 [2] があるが、本研究では類似度として Dice 係数を用いた。また、通常は精度を重視して、類似度が最大のものを候補とするが、本研究ではカバレッジを重視したため、一つの日本語表現に対して、Dice 係数 0.7 以上で上位 3 位までの英語表現を対訳表現として抽出した。なお、日本語では平仮名を含めた文字  $N$  グラム、英語では単語  $N$  グラムをそれぞれ用い、ともに  $N = 25$  を最大値とした。その結果、名詞など単語に対する訳語だけでなく、〈なお従前の例による / it shall be handled as heretofore〉のように、定型表現に対する対訳表現も獲得することができた。

### 3.3 人手による対訳表現の選抜

自動抽出された対訳表現には適切でないものも含まれるので、法律専門家が 2 度にわたり、除去・修正を行った。その結果を表 1 に示す。人手による第 1 次選

表 1: 対訳表現の獲得結果

法令 文書 セット	対応 文数	計算機に よる自動 抽出	対訳表現					
			人手による選抜					
			Bilingual KWIC 未使用		Bilingual KWIC 使用			
			第 1 次	第 2 次	第 3 次	第 4 次		
A	26,257	49,745	15,399					
B	13,303	17,009	6,776					
A+B	39,560		対訳表現数 (重複除去)	21,354	4,954	3,329	4,037	(2,414)
			日本語見出し語数	14,474	2,555	2,246	3,329	(2,098)
			対訳表現数平均	1.4	1.9	1.5	1.2	(1.2)

抜では、見出し語が日本語として不適切なものや、対訳として誤っているものを取り除いた。なお、法令文は各府省から数回に渡って提供されたため、自動抽出と第 1 次選抜も 2 回に分けて行った。表 1 では、それぞれの結果を A,B として示している。

第 2 次選抜では、法令翻訳における標準辞書という観点から登録するのに相応しい語を残した。その結果、日本語見出し語 2,555 個、対訳表現 4,954 個を獲得した。この時点における日本語見出し語 1 語あたりの対訳表現数は 1.9 であった。

## 4 GUI による対訳辞書作成支援

前節の方法で獲得した対訳表現について、法律専門家が対訳コーパスを参照しながら吟味するとともに、用例や使い分け・コメントなどを付加し、対訳辞書を作成した。その支援のために、図 2 に示す対訳辞書作成支援システム Bilingual KWIC[3][4] を用いた。

### 4.1 Bilingual KWIC

Bilingual KWIC は対訳コーパスからの対訳自動抽出を可視化したソフトウェアである。図 2 の画面左上のキーワード入力欄にキーワードを入力し、その横の [Search] ボタンを押すと、左側に原文、右側に対応付けられた対訳文を表示する。その際、原文ではキーワードを中心に、また、対訳文では自動的に推定したその対訳表現を中心に、それぞれ KWIC 形式で表示する。また、注目する文をマウスでクリックすると、その文全体が下側に表示される。

Bilingual KWIC は対訳表現を自動的に推定するため、対訳辞書は不要であるが、既存の辞書と併用することも可能である。入力されたキーワードの訳語が辞書に登録されている場合、そのキーワードと訳語は緑

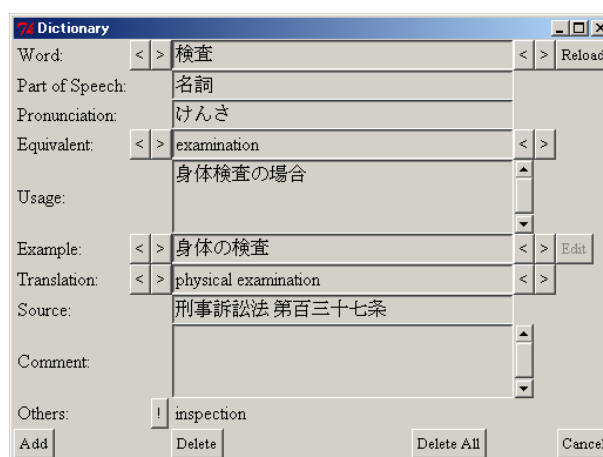


図 1: Bilingual KWIC における辞書登録

色で表示される。その後、原文にキーワードを含むが、対訳文に辞書にある訳語を含まない文を集め、再帰的に対訳自動抽出を行い、その訳語を青色で表示する。

図 2 では、「検査」をキーワードとして入力したところ、辞書に登録されている“inspection”以外に、“physical examination”、“examining”、“be examined”が対訳表現として抽出されている。ここで、“physical examination”は「検査」ではなく「身体検査」の訳語であるが、原文と対訳文が KWIC 表示されていることより、そうした間違いの修正も容易である。

さらに、Bilingual KWIC では、辞書に登録したい単語を右クリックすることにより、図 1 に示す辞書登録が行える。見出し語 [Word]、対訳表現 [Equivalent]、用例 [Example] および、その訳 [Translation] の欄はコーパス中の文の一部を抽出して登録できる。それに加えて、訳し分けの情報やコメントなども自由に記述して辞書に登録できる。見出し語に対して、他の訳語が登録されている場合は [Others] 欄に表示される。

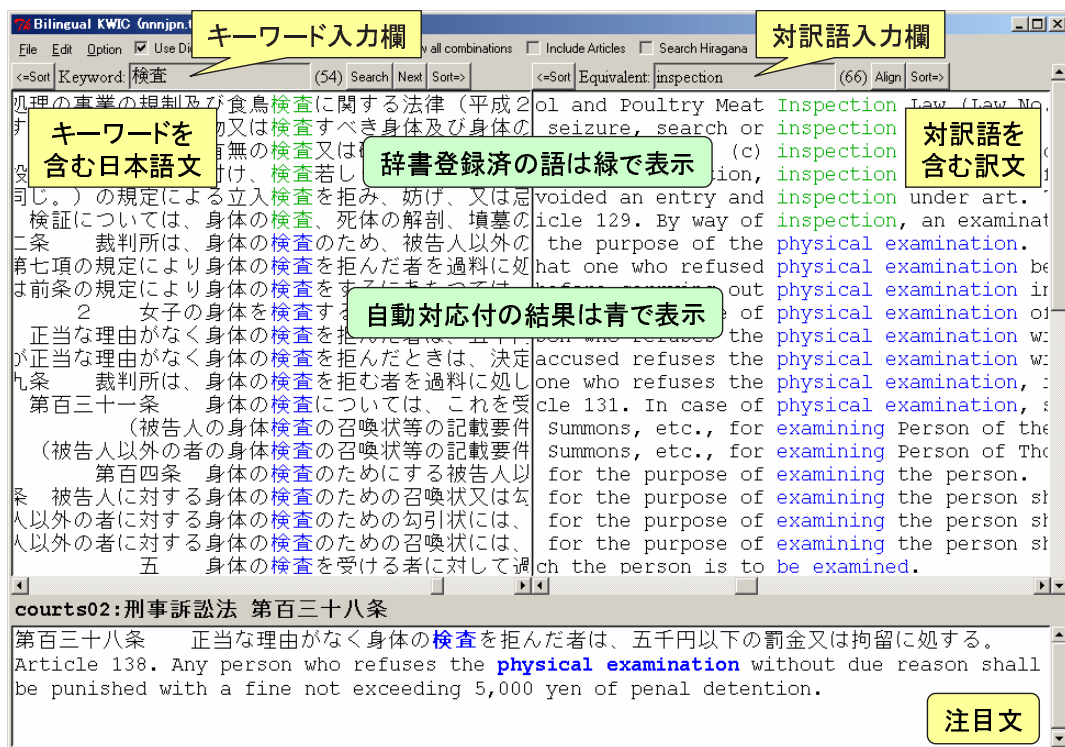


図 2: Bilingual KWIC の概観

## 4.2 暫定版 標準対訳辞書

この Bilingual KWIC をインストールした PC を各府省に貸し出し、自動抽出の結果の修正およびコメントなどの付加を、法律専門家に行ってもらった。その結果、表 1 の第 3 次の欄に示すように、2,246 語の見出し語からなる標準対訳辞書の暫定版ができた。なお、1 語あたりの対訳表現数は 1.5 となった。しかし、約 2,000 語では法令翻訳用の辞書としては不十分である。収録語数が少ない原因は二つあると考えられる。

一つは、対訳自動抽出における見出し語の選定にある。今回は、日本語の見出し語もコーパス中から自動的に抽出し、対訳表現との類似度が閾値以上のものだけを残した。そのため、複数の対訳表現をもち、そのいずれに対しても閾値を超えなかった日本語見出し語は登録されなかった。この点に関しては、日本語の見出し語を既存の法令用語集などから収集し、それを併用することによってカバレッジを上げるべきであった。

もう一つの原因は、対訳コーパスの分野の偏りである。現在、有効な法令は約 7,100 本あり、今回使用した法令 153 本はその 1 割にも満たない。そのため、重要な法律用語でありながらも今回のコーパス中に出現しないものがあり、そうした語は今回の手法では辞書

登録できなかった。

また、「この法律は、…から施行する」のように、間に他の語を含む定型表現についても登録できなかった。

そこで、専門家が既存の法律用語の対訳辞書と Bilingual KWIC を併用し、不足する用語を辞書に追加した。その結果、現在、表 1 の第 4 次の欄に示す 3,329 語が標準対訳辞書に登録されている。なお、括弧内の数値は人手による追加分を除いた結果である。また、各対訳表現についても学識経験者による検討がさらに加えられ、対訳表現数の平均は 1.2 に減少している。

## 5 訳語統一支援

標準対訳辞書が完成した後は、その辞書に準拠して法令を翻訳することになる。しかし、既に法令の翻訳がある場合には、最初から翻訳をやり直すよりも、既にある翻訳を標準対訳辞書に合わせて修正した方がコストが少なく済む。そこで、標準対訳辞書に準拠した訳語統一を支援することを目的として、図 3 に示す訳文添削システムを開発している。

この添削システムは、(1) 新たに高機能な添削システムを作成するコストが削減できる、(2) 実際の作業者が習熟しているシステムである、との理由により、MS-Word を利用している。

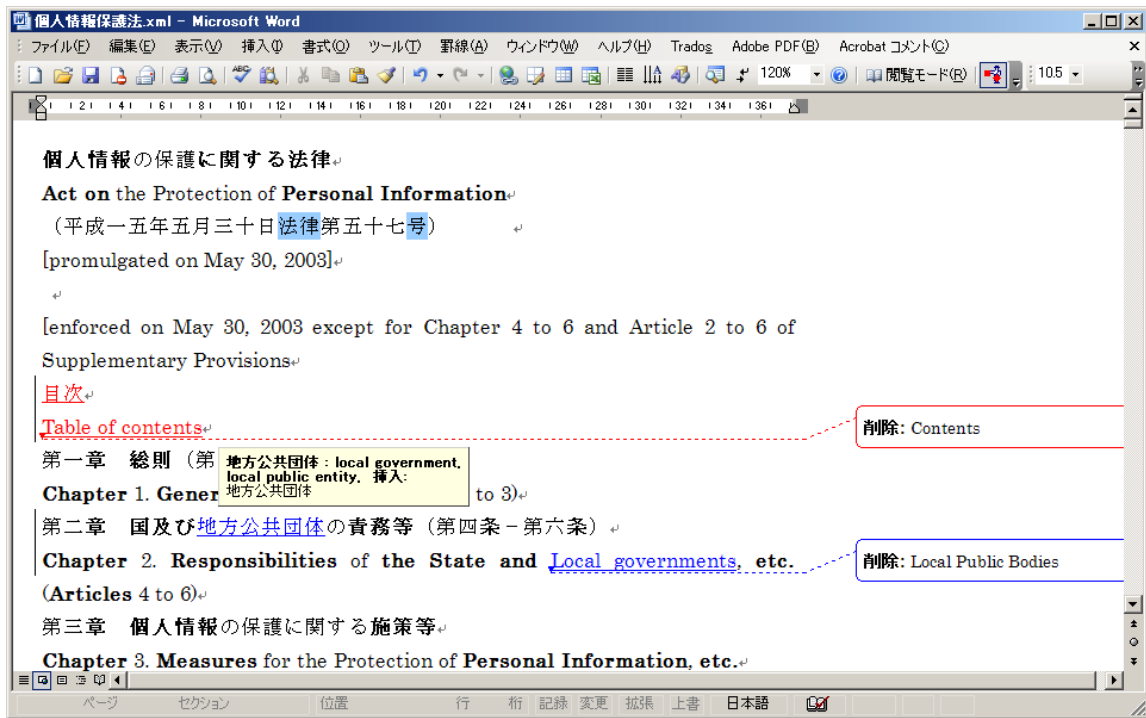


図 3: 訳文添削システムの概観

本添削システムでは、日本語の原文と英語の対訳文を交互に表示する。原文中に出現した見出し語が、対応する対訳文中において標準対訳辞書にある標準訳に翻訳されている場合、両者を太字で表示する。原文中に見出し語があるが、対訳文中に標準訳がない場合、対訳文中における対訳表現を推定し、それを標準訳で置換する。図 3 の例では、原文中 12 行目の「地方公共団体」は、最初は“Local Public Bodies”と翻訳されていたが、“Local governments”に置換されている。

一方、対訳表現が推定できなかった場合には、原文中に見出し語部分の背景を青色で表示する。図 3 の例では、3 行目の「法律」「号」に対し、4 行目の英文に対応する訳語がないため、背景が青色で表示されている。

なお、いずれの場合も見出し語や添削箇所にもマウスカーソルを近づけると、標準対訳辞書の内容がポップアップで表示される。これにより、ユーザは辞書を手軽に参照して、翻訳文を添削することができる。

なお、本システムの技術的詳細については、稿を改めて述べる予定である。

## 6 おわりに

本稿では、対訳自動抽出の技術を実際の法令翻訳辞書の作成に利用した事例について紹介した。法令英訳

用の標準対訳辞書第 1 版と、第 5 節のシステムを用いて作成した英訳法令 14 本は、政府から今年度中に公開される予定である。また、来年度以降、その他の法令も順次、翻訳・公開される予定である。

## 謝辞

本研究の推進にあたり、内閣官房司法制度改革推進室の山本拓参事官補佐と中川明子参事官補佐に協力いただきました。記して感謝します。

## 参考文献

- [1] 外山 勝彦, 小川 泰弘, 松浦 好治: 日本法令翻訳システムの構想, ジュリスト, No.1281, pp.2-5 (2004).
- [2] Yuji Mastumoto and Takehito Utsuro: Lexical Knowledge Acquisition, *Handbook of Natural Language Processing*, Dale, R., Moisl, H., and Somers, H.(Eds.), pp.563-610, Marcel Dekker (2000).
- [3] 小川 泰弘, 西森 寛敏, 外山 勝彦: Bilingual KWIC - 対訳抽出の可視化による翻訳支援, 言語処理学会第 11 回年次大会講演論文集, pp.811-814 (2005).
- [4] OGAWA Yasuhiro and TOYAMA Katsuhiko: Bilingual KWIC - GUI Support Tool for Bilingual Dictionary Compilation-, *Proceeding of The 6th Symposium on Natural Language Processing (SNLP 2005) in concurrent with PAPILLON 2005*, Vol.2, pp.77-84 (2005).